2016-7-A3-4

DNNによる音声認識を用いた感情音声の声質変換の検討

笹田拓臣[†1]　相澤佳孝[†1]　小坂哲夫[†1]

声質変換の方法として，音声認識の結果として得られた音素系列を用いて音声合成を行う方法が提案されている．この方法で感情音声を対象とした場合，認識精度の低下が問題となる．このため，変換音声の品質が低下してしまう．この問題に対して本研究では，感情適応を用いて音声認識を行う方法を提案する．これにより，感情音声の認識精度の向上による変換音声の品質の改善を目指す．

A Study on Voice Conversion of Emotional Speech using Speech Recognition by DNN

Takumi Sasada,[†1] Yoshitaka Aizawa[†1]
and Tetsuo Kosaka[†1]

The voice conversion method where speech synthesis is conducted using a phoneme sequence obtained as a result of speech recognition has been proposed. However, it has a problem with declining recognition performance when targeting emotional speech. This will cause the degradation of the converted speech quality. To solve this problem, we propose a use of emotion adaptation for speech recognition. In this way, we aim to improve the quality of converted speech by improving recognition accuracy of emotional speech.

1. はじめに

任意の内容から音声を合成する方式は規則合成方式という．これは合成したい音声に対応した音素や音節等の規則を入力し，それらの情報に基づいて音声の合成を行う．特に，発話内容を記したテキストから対応する音声を合成する技術はテキスト音声合成（Text To Speech: TTS）といい，近年その技術が急速に発達した．その背景として，音声合成用の大規模な音声コーパスの整備や，コンピュータの計算能力の向上，そしてコーパスベースと呼ばれる大量のデータを用いた自動学習および音声単位選択に基づいた手法の発達がある．現在では隠れマルコフモデル（Hidden Markov Model: HMM）を用いた音声合成手法[1)]が注目されており，その数学的な取り扱い易さと柔軟性から様々な機関にて盛んに研究が行われている．

一方で，入力された音声を特定話者の音声へ変換する「声質変換」と呼ばれる技術がある．従来のガウス混合モデル（Gaussian Mixture Model: GMM）に基づいた手法では，モデル学習に発話内容が入力音声話者と変換先音声話者で対応している音声データが必要である等の問題があったが，近年では対応している音声データを必要としない，非パラレル声質変換も提案されている．例えば，HMM 音素認識と HMM 音声合成を利用した HMM 声質変換[2)]はこれに該当する．一般的に声質変換では話者性も含めた変換を行うため，入力音声の韻律情報に含まれている話者性も変換先話者の話者性に変換される．これは変換先話者の再現性は高まるものの,「ユーザが思い通りの発話表現を持つ音声の合成を可能にする」という要望にはそぐわない恐れがある．このような問題に対し，HMM 音声合成の枠組みでも入力音声による韻律情報の制御機能を持つ音声合成システムが提案されている[3)]．

また近年では，ディープニューラルネットワーク（Deep Neural Network: DNN）を用いた音声認識手法が提案されている．この手法は音声認識において高い成果を上げており[4)]，Google や Microsoft などが提供する音声認識サービスも DNN ベースの物に置き換わっている．また，DNN-HMM を用いた声質変換手法も提案されている[5)]．音声認識・合成による声質変換手法においても認識精度が高いとされる DNN-HMM を用いることで，合成音声品質が向上することが期待される．

現状では，感情が含まれていない音声であれば音声認識は高い精度を得ている．しかし，感情が含まれている音声の認識は未だ難しく，精度が低いのが現状である．このため，音声認識・合成による声質変換では，変換音声の品質が低下してしまうという問題がある．この問題を解決する手段として，DNN を用いる方法が考えられる．この DNN を用いた感情音声認識手法は奈良先端大より提案されている[6)]．しかし，この手法では言語モデルがタスククローズであり，汎用の発話内容の認識ができないという問題がある．そこで本研究では，大規模なコーパスから学習した言語モデルを使って認識を行う．さらに，感情音声を使用して音声認識用の音響モデル適応を行う手法を提案する．これにより，感情音声の認識精度の向上による変換音声の品質の改善を目指す．使用する感情音声は，感情評定値付きオン

†1 山形大学
Yamagata University

ラインゲーム音声チャットコーパス（Online Gaming Voice chat Corpus with emotional label: OGVC）[7] を使用する．

2. 音声入力による声質変換システム

2.1 声質変換システムの概要

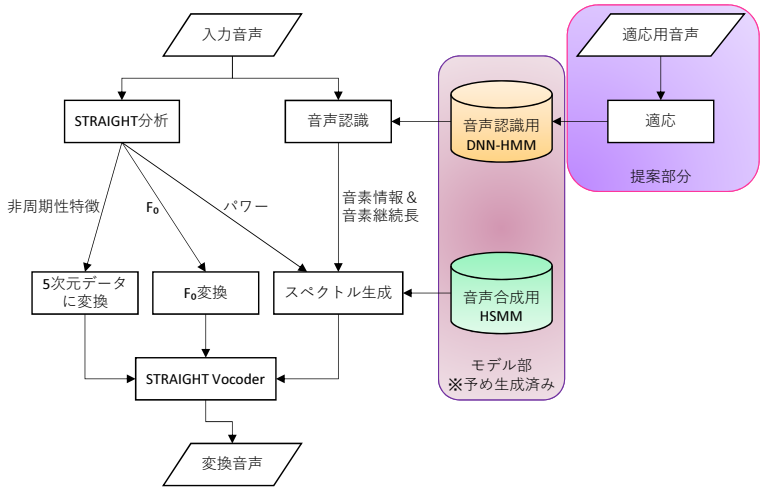

図 1 声質変換システム
Fig. 1 Voice Conversion System

声質変換システムの具体的な構造を図 1 に示す．このシステムでは入力音声から必要な韻律情報と音素情報を抽出して音声を合成・出力する．一般的な声質変換では韻律も変換先に合わせるのが通常であるが，本研究では韻律を入力音声で制御することが目的であるため，韻律情報は入力音声から抽出する．音素情報は音声認識により生成し，認識用のモデルは DNN-HMM を，合成用のモデルは HSMM（Hidden Semi-Markov Model）を使用する．認識用のモデルは従来 GMM-HMM が用いられていたが，先に述べた通り近年 DNN を用いた音声認識手法が高い成果を上げていることや，従来の GMM-HMM を用いた音声認識手法よりも DNN-HMM を用いた音声認識手法の方が認識精度が改善したという報告[8] もあることから，本研究でも認識用モデルは DNN-HMM を使用する．なお，これらのモデルは予め学習を行い用意しておく必要がある．以下，このシステムによる声質変換処理の流れを説明する．

まず，入力音声の音声認識を行い，音素情報とそれに対応する音素継続長を抽出する．分析された音素継続長の情報を用いて，変換音声話者のメルケプストラム系列を音声合成用 HSMM を用いて生成する．

更に，入力音声から高品質音声合成分析システム STRAIGHT[9] を用いて，ピッチ・非周期性特徴・パワーの 3 つを分析する．ピッチと非周期性特徴は直接 STRAIGHT から出力されるが，パワーは STRAIGHT スペクトルに対しメルケプストラム分析を行い，その 0 次の成分を合成時に用いることでパワー情報を反映させる．

音素継続長を用いて生成されたメルケプストラム係数列と，パワー情報から生成した利得を合わせ，スペクトルに変換する．また，ピッチは変換話者の対数平均に調整され，非周期性特徴は 5 帯域で分割，帯域ごとの平均値を得る．パワー情報も反映したスペクトル系列とピッチ，非周期性特徴を STRAIGHT の Vocoder（音声合成器）を用いて，目的の音声を合成する．

文献[2] では入力音声から分析したピッチを量子化し，コンテキストとして用いている．結果として最終的に出力される変換音声のピッチも，音声合成用モデルから出力されたものになるが，本研究では文献[3] と同様に，入力音声から分析したピッチを線型変換させて直接用いる．

また，DNN-HMM を用いた非パラレル声質変換手法が文献[5] にて提案されており，その手法では継続長の指標に音素状態を用いているが，本研究では音素を使用する．

次に，本研究での提案法について説明する．DNN-HMM を用いることで高い認識精度が得られているが，これは感情が含まれていない音声の場合である．感情が含まれている音声の場合，認識が困難であり精度が低いというのが現状である．以上の理由から本研究では，上記声質変換システムの音声認識部の HMM に対し適応を用いることで，認識精度が向上する方法を提案する．

2.2 モデル適応手法

DNN-HMM の適応手法としては，学習時と同様に Fine-tuning と呼ばれる方法を用いる．これはフレームごとに状態番号を与え，確率的勾配降下法（Stochastic Gradient Descent: SGD）による誤差逆伝搬法による教師付き学習を行う．損失関数にはクロスエントロピーを用いる．認識時にはベイズ則に基づくスケーリングを行って出力確率を求め，HMM を用いた確率計算を行う．適応のパラメータとして遷移確率の更新も考えられるが，本研究で

は DNN のみのパラメータ更新を行う．DNN の適応を行う場合，過学習が問題となる．この問題に対処する方法として，モメンタムや正則化などを用いる手法が検討されている[10]．基本的にはモデルの自由度を制限することにより過学習を抑制する．また，Dropout[11] と呼ばれる学習時の各反復において，一部のノードをランダムに取り除いて学習する方法も過学習に有効であると考えられる．以上に挙げた手法のうち，本研究では L2 正則化を利用する．

2.3 事後確率の補正手法

DNN-HMM における出力確率計算には，次式に示すベイズ則を用いる．

$$p(x|s_i) = \frac{p(s_i|x)p(x)}{p(s_i)} \tag{1}$$

ここで，$p(s_i|x)$ は DNN から得られる出力，$p(x)$ は入力特徴量の生起確率，$p(s_i)$ は状態生起確率を表す．この $p(s_i)$ を理論通りに与えた場合，無音（sil）に関係する状態の生起確率が極端に大きいため，システムの認識性能が十分に向上しないことが分かっている．

そこで，この状態生起確率の偏りを解消するために，DNN の学習時の Forward 計算に用いられる音素カウントファイル（各状態のカウント値を記録したもの）に変更を加えることにより $p(s_i)$ を修正し，事後確率を補正する．この変更とは，状態 s_i の生起確率 $p(s_i)$ に対して閾値（上限値）θ を設定し，上限を超えた値を θ に置き換える，というものである．この際 θ は，制限率 $\alpha(0 \leq \alpha \leq 1)$ を指定した上で，次式に基づいて決定する．

$$\alpha = \frac{\sum_{i \in D}\{p(s_i) - \theta\}}{\sum_{i=1}^{I} p(s_i)} \tag{2}$$

ここで，i は状態番号，I は状態数を表し，$p(s_i) > \theta$ を満たす i の集合を D とする．$p(s_i)$ が大きい順に i を並び替えた場合の模式図を図 2 に示す．

3. 基本的な実験条件

3.1 音声認識の実験条件

入力音声の音素情報と音素継続長は，音声認識用の DNN-HMM を用いた音声認識によって抽出される．本研究における音声認識の実験条件を表 1 に示す．認識には研究室独自の 2 パスデコーダを用いる．本研究で用いる認識システムは，第 1 パスで triphone と bigram を用いてビームサーチを行って単語グラフを作成し，第 2 パスでは生成した単語グラフを trigram でリスコアし認識結果を得る構成となっている．また一般的に DNN-HMM を用い

表 1 音声認識の実験条件
Table 1 Experimental Conditions of Speech Recognition

音響モデル	
コーパス	CSJ 学会講演および模擬講演 963 講演
モデル形式	不特定話者モデル 3003 状態
分析条件	基本周波数：16kHz，量子化：16bit，分析窓：ハミング窓， フレーム長：25msec，フレーム周期：8msec， 高域強調：0.97，Δ窓幅：2 フレーム
特徴ベクトル	フィルタバンク（25 次），Δ，Δ^2 計 75 次元
言語モデル	
学習データ	CSJ 学会講演および模擬講演 2702 講演 (総単語数：約 668 万語)
カットオフ	unigram:1，bigram:1，trigram:3
DNN 構造	
入力フレーム	11 フレーム
入力層	825 ノード
中間層	2048 ノード × 7 層
出力層	3003 ノード，HMM の状態確率として利用
適応条件	
適応データ	OGVC 演技音声の 4 話者（男女各 2 名，各話者 112 発話）
学習係数	0.0001
ミニバッチサイズ	2048
エポック数	15
モメンタム	0
L2 正則化係数	0.0002
音素カウント制限率	0.1
認識条件	
第一パス	言語重み：10，挿入ペナルティ：-8
第二パス	言語重み：14，挿入ペナルティ：-8
ビーム幅	単語内：200，単語間：120，仮説数：2400
評価条件	
評価データ	OGVC 演技音声の 4 話者のうち感情強度 3(強) の音声 （男女各 2 名，各話者 112 発話）

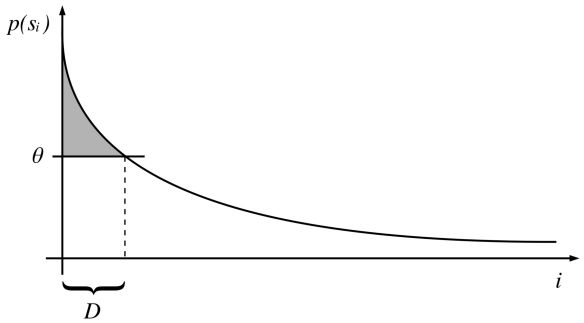

図 2 事後確率の補正の模式図
Fig. 2 Schematic Diagram of Correction of Posterior Probability

た音声認識では，複数フレームの特徴ベクトルをひとまとめにしたセグメント特徴量が用いられる．本研究でも 11 フレームの特徴を入力とする．隠れ層の総数については，日本語話し言葉コーパス（Corpus of Spontaneous Japanese: CSJ）の学習データ量では 5〜7 層程度で飽和することが示されているため[12)]，本研究では 7 層とした．また，ノード数は 512〜2048 程度が使用されるが，本研究では 2048 とした．出力層は，ハイブリッド型の場合，認識に用いる HMM の総状態数に揃える必要がある．本研究では triphone を用い 3003 ノードとした．なお，本実験では，適応・評価ともに OGVC の演技音声を利用する．OGVC 演技音声には話者 4 人（FOY，FYN，MOY，MTY）と感情強度 4 段階（0:平静，1:弱，2:中，3:強）の組み合わせのサブセットが収録されている．また各サブセットには 166 発話が収録されているが，モーラ数が 5 以下の文を排除した 112 発話を使用した．

3.2 音声合成の実験条件

システムに入力された音声は，STRAIGHT 分析によってピッチ・非周期性特徴・パワーの 3 つの韻律情報が分析される．また音声合成は，HTS を用いて生成した音声合成用 HSMM によってメルケプストラムを生成し，このメルケプストラムと他方で分析された韻律情報を用いて STRAIGHT Vocoder による音声合成を行う．音声合成の実験条件を表 2 に示す．

4. 音響モデル適応による音声認識実験

4.1 実 験 条 件

本実験では，適応データを様々に変更して適応・認識実験を行い，どの適応法が最も認識精度が改善され，かつ実用的であるか調査した．評価音声は OGVC の感情強度 3(強) の音声のみを対象とした．適応法に関しては，次の 2 つの仮定に基づいて 5 種類を設定した．その設定した適応の種類を表 3 に示す．これは表 1 の適応条件の適応データの部分のことを示している．

話者適応 話者ごとに感情表現は異なると仮定．単一話者のデータで適応する．つまり，評価データと同様のデータを適応データとして用いる．

感情適応 話者による感情表現には大きな差がないと仮定．複数話者による同一感情のデータを用いて適応する．感情適応 (3) は OGVC 演技音声の話者 4 人のうち評価話者以外の 3 人の感情強度 3(強) の正解を適応データとして用いる．さらに感情の強度によって感情表現は異なると考えられるので，強度別に適応する方法も行う．感情適応 (1〜3) は評価話者以外の 3 人の感情強度 1(弱)〜3(強) の計 9 種の正解を適応データとして用いる．

また，この適応を用いた音声認識実験条件は基本的には表 1 に同じであるが，適応条件の

表 2 音声合成の実験条件
Table 2 Experimental Conditions of Speech Synthesis

合成用音響モデル	
学習音声	ATR デジタル音声データベース セット B
コーパス	MYI，FKN 話者 連続発声（SD） A〜I セット 450 文
モデル構造	1 音素毎に 5 状態 1 混合のトライフォン HSMM
総状態数	FKN 話者：MCEP464 状態，継続長 125 状態 MYI 話者：MCEP486 状態，継続長 124 状態
STRAIGHT 分析条件	
標本化周波数	16000Hz
量子化ビット数	16bits
分析周期	5msec
FFT 長	2048
F0 分析最小値	50Hz
F0 分析最大値	670Hz
特徴ベクトル	0〜39 次のメルケプストラム及びその Δ，Δ^2
メルケプストラム分析条件	
メル尺度係数	0.42
ガンマ係数	0

適応データの部分のみが異なる．その部分を適応法ごとに表 4 に示す．

表 3 適応の種類
Table 3 Types of Adaptation

適応法	適応データとその目的
適応なし （ベースライン）	適応せずに認識（従来法と同様） → ベースラインとして比較
正解学習 （目標値）	認識結果の代わりに正解を与え学習 → 認識精度の目標値として比較
話者適応	評価データ (同一話者) を用いて教師なし適応 → 話者性と感情の両方に着目
感情適応 (1〜3)	評価話者以外の感情強度 1(弱)〜3(強) の正解を用いて教師付き適応 → 感情強度のみに着目
感情適応 (3)	評価話者以外の感情強度 3(強) の正解を用いて教師付き適応 → 感情強度が強い音声のみに着目

表 4 適応法ごとの音声認識実験条件
Table 4 Experimental Condition of Speech Recognition by Adaptation Method

適応データ	
ベースライン	なし
目標値	評価データの正解 (各話者 112 発話)
話者適応	評価データの認識結果 (各話者 112 発話)
感情適応 (1〜3)	評価話者以外の感情強度 1(弱)〜3(強) の正解 (各話者につき，112 発話 × 3 話者 × 3 強度 = 1008 発話)
感情適応 (3)	評価話者以外の感情強度 3(強) の正解 (各話者につき，112 発話 × 3 話者 = 336 発話)
評価データ	
OGVC 演技音声の 4 話者のうち感情強度 3(強) の音声 (男女各 2 名，各話者 112 発話)	

4.2 実験結果および考察

単語認識実験結果を図 3 に，音素認識実験結果を図 4 に示す．図 3，4 より，話者適応の場合の認識結果は適応なしの場合よりも認識率が向上していることが分かる．これより，話

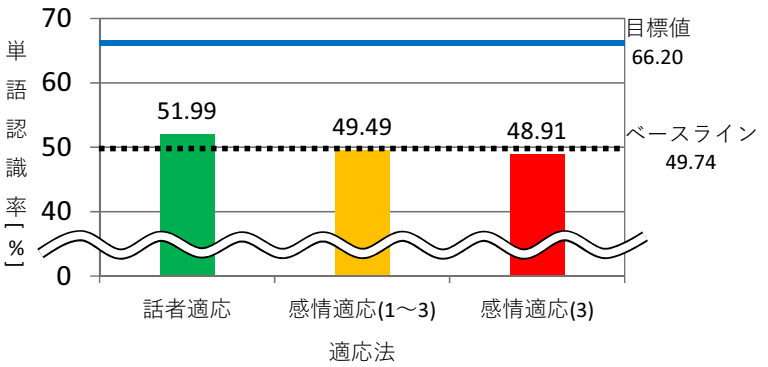

図 3 単語認識実験結果
Fig. 3 Word Recognition Results

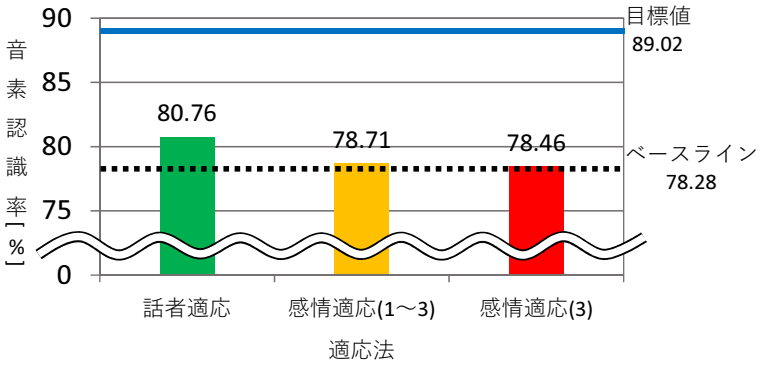

図 4 音素認識実験結果
Fig. 4 Phoneme Recognition Results

者適応は認識精度の向上に効果があると考えることができる．目標値の場合の認識結果は最も認識率が高い結果となっているが，これは評価データの正解を用いた適応であるため，実用的ではない．感情適応 (1〜3)・感情適応 (3) の場合の認識結果は，単語認識率は適応なしよりも低下してしまったが，音素認識率は適応なしよりも向上している．しかしその差は決して大きな値ではないので，認識精度の向上にはあまり効果がない適応法であると考えられる．

以上より，誤り率・実用性を考えた結果，話者適応が最も有用性のある適応法であると考えることができ，その適応法を用いた場合の認識精度の向上について今後検討する必要があると考えられる．

4.3 事後確率補正の予備実験

前節で最も有用性のある適応であると結論付けた話者適応に焦点を当て，音素カウント制限率を 0.00〜0.50 の範囲で 0.05 刻みで設定し，適応・認識実験を行った．但し，制限率 0.00 は“制限なし”と同様である．この実験の結果は，制限率が 0.30 の場合が僅かであるが他の制限率よりも認識精度の改善が見られた．

5. 変換音声の主観評価実験

5.1 実 験 条 件

本実験では，まず表 2 の条件で変換音声を生成し，その生成した変換音声の品質を主観評価実験により調査した．生成した変換音声の種類は,「ベースライン」「話者適応」「目標値」の 3 つである．各々のデータについては表 3 と同様であり，音素カウント制限率については，ベースラインは適応なしのため制限なし，話者適応は制限率 0.30，目標値は制限率 0.10 とした．話者適応の場合の制限率を 0.30 としたのは，4.3 節の実験結果を踏まえたためである．また，本実験で使用する評価項目は次の 2 項目である．

了解度　変換音声の発話内容がはっきりと聞き取れるか．ここでは，入力音声の発話内容と変換音声の発話内容が聴取して一致しているかどうか．

声質　ノイズや耳障りな音が無く，人間の音声らしく聴きやすいか．

上記の主観評価項目のうち，了解度に関しては変換音声の発話内容を聴き取りそれを書き起こしたものを音素列に変換し，入力音声の音素列と比較する．声質に関しては変換音声に対して 1〜5 までの 5 段階で絶対評価を行う MOS（Mean Opinion Score）評価を行う．被験者は OGVC にあまり慣れていない 20 代学生 10 名で，1 人当たり 48 文を評価した．この 48 文の中に変換音声 3 種類（ベースライン，話者適応，目標値）が各 16 文ずつランダムに含まれている．なお，評価に使用する音声は全て異性間変換した音声のみを使用し，同性間変換した音声は用いていない．評価の際の指標は表 5 のように設定した．

表 5　評価指標
Table 5　Evaluation Index

評点	カテゴリ
5	非常に良い（Excellent）
4	良い（Good）
3	普通（Fair）
2	悪い（Poor）
1	非常に悪い（Bad）

5.2 実験結果および考察

主観評価実験による了解度の結果を図 5 に，声質評価の結果を図 6 に示す．図 6 のエラーバーは 95%信頼区間である．

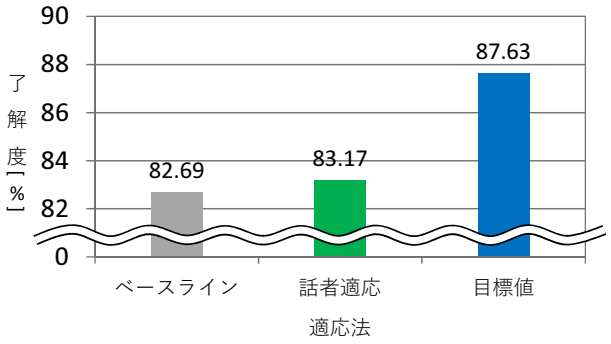

図 5　主観評価実験による了解度結果
Fig. 5　Intelligibility Results by Subjective Evaluation

図 5 より，了解度はベースラインよりも話者適応の方が数値としては高くなっているが，その差は決して大きな値ではなく，ほぼ同等であると言える．目標値の了解度はその 2 つよりも高い数値を示していることから，了解度の向上には認識精度の改善が必要であり，そのためには適応の精度を上げる必要があると考えられる．

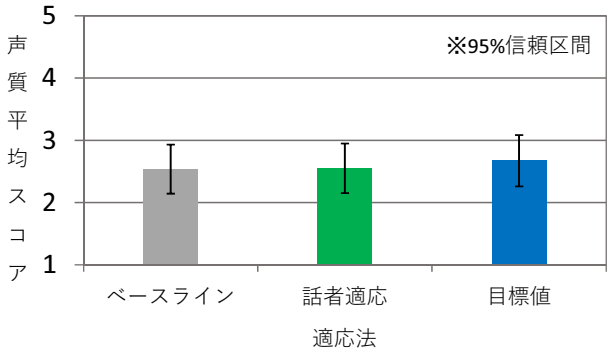

図 6 主観評価実験による声質評価結果
Fig. 6 Voice Quality Results by Subjective Evaluation

図 6 より，声質に関しては 3 つ全ての方法において大差のない平均スコアになっている．また，95%信頼区間のエラーバーは全て重なっており，有意差は見られなかった．これより，了解度と声質の相関関係は小さいことが分かるので，了解度とはまた別に声質に関しても今後改善していく必要があると考えられる．

生成した変換音声が従来法よりも改善した例として，発話内容が「準備時間なんだよ」の「準備」の部分を拡大した変換音声のスペクトルを図 7，8 に示す．図 7 が従来法，図 8 が提案法のスペクトルである．この 2 つの図より，従来法では「z e N b ee」と認識されていたが，本研究の提案法では「j u N b i」と正解の発話と同様の認識結果となっていることが分かる．

さらに主観評価の際に被験者から頂いた意見として，「子音が他の子音に変わっていることが多かった」や「ノイズや雑音が多かった」，「全体的に聴き取りづらかった」，「何を言っているのか分からない音声も多かった」等があった．これらはいずれも了解度・声質に関係することであるので，これらを改善することができれば，了解度・声質の向上が期待できる．

6. 結論および今後の課題

本研究の目的は，声質変換システムの音声認識部に音響モデル適応を行うことにより，感情音声の認識精度の改善による変換音声の品質の向上であった．単語認識実験結果および音素認識実験結果より，話者適応が最も認識率が向上していることから，認識精度の改善には感情よりも話者の個人性の方が重要であることが分かった．一方，主観評価実験結果より，

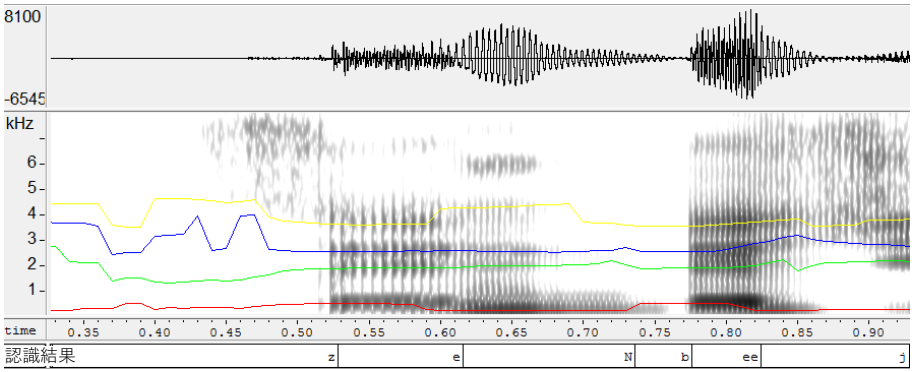

図 7 従来法による変換音声の「準備」の部分のスペクトル
Fig. 7 Spectrum of the Converted Speech "Junbi" by the Conventional Method

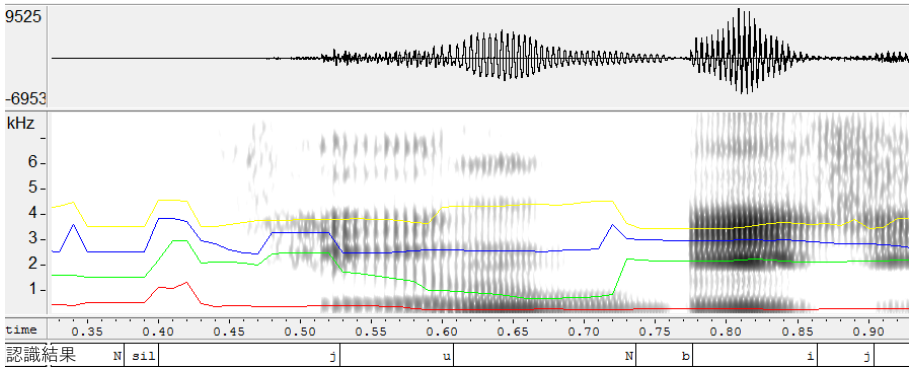

図 8 提案法による変換音声の「準備」の部分のスペクトル
Fig. 8 Spectrum of the Converted Speech "Junbi" by the Proposed Method

了解度および声質はベースラインからほとんど向上が見られなかっただけでなく，了解度と声質の相関関係は小さいことも分かった．

認識精度は改善したものの了解度・声質はほとんど向上が見られなかったという結果となったが，主観評価実験結果の目標値に注目すると，了解度は向上が見られていることから，認識精度が上がれば了解度も向上すると考えられる．よって，適応の性能を上げれば認識精度も改善され，了解度も向上することが期待できる．今後は適応データをより多く用意し利用できれば，適応の性能向上に繋がると考えられるので，その検討をする余地がある．また声質に関しても，ピッチと音素の種類の有声/無声不一致により，例えばピッチが有声で音素の種類が無声の場合はノイズになってしまい声質を低下させている原因になっていると考えられる．この対処法として，その区間の前後の音素の継続長の変更による補正処理等の検討をする必要がある．

参 考 文 献

1) 吉村貴克，徳田恵一，益子貴史，小林隆夫，“HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化”，電子情報通信学会論文誌，vol.J83-D-II(11)，pp.2099-2107，2000.
2) 太田悠平，能勢隆，小林隆夫，“量子化 F0 コンテキストを用いた HMM に基づく不特定話者声質変換の検討”，日本音響学会講演論文集，1-7-22，pp.327-328，2010.
3) 西垣友理，高道慎之介，戸田智基，Graham Neubig，Sakriani Sakti，中村哲，“音声入力による韻律制御機能を有する HMM 音声合成システム”，日本音響学会講演論文集，3-6-11，pp.343-344，2014.
4) 神田直之，武田龍，大淵康成，“Deep Neural Network に基づく日本語音声認識の基礎評価”，情報処理学会研究報告，vol.2013-SLP-97(8)，pp.1-6，2013.
5) Minghui Dong et al.，“Mapping Frames with DNN-HMM Recognizer for Nonparallel Voice Conversion”，APSIPA ASC，pp.488-494，2015.
6) 向原康平，サクティ・サクリアニ，吉野幸一郎，ニュービッグ・グラム，中村哲，“ボトルネック特徴量を用いた感情音声認識の検討”，日本音響学会講演論文集，2-1-7，pp.43-44，2016.
7) 有本泰子，河津宏美，“音声チャットを利用したオンラインゲーム感情音声コーパス”，日本音響学会講演論文集，1-P-46a，pp.385-388，2013.
8) 相澤佳孝，中川由暁，小坂哲夫，加藤正治，“HMM 認識・合成による感情音声の声質変換の性能向上”，日本音響学会講演論文集，3-Q-32，pp.269-272，2016.
9) “STRAIGHT information”，http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_j.html.
10) H.Liao，“Speaker adaptation of context dependent deep neural networks”，Proc. of ICASSP2013，pp.7947-7951，2013.
11) G.E. Dahl，T.N. Sainath and G.E. Hinton，“Improving deep neural networks for LVCSR using rectified linearunits and dropout”，Proc. of ICASSP2013，pp.8609-8613，2013.
12) 三村正人，河原達也，“CSJ を用いた日本語講演音声認識への DNN-HMM の適用と話者適応の検討”，情報処理学会研究報告，2013-SLP-97(9)，pp.1-6，2013.